



National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

UNIVERSITY OF ALBERTA

THE EFFECT OF HIGHER ORDER LATENT SPACES
ON
THE ROBUSTNESS OF UNIDIMENSIONAL NONLINEAR ITEM RESPONSE MODELS

by

ALBERT LEO BEAULNE

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF EDUCATION

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL, 1990



National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-64875-9

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR

ALBERT LEO BEAULNE

TITLE OF THESIS

THE EFFECT OF HIGHER ORDER LATENT SPACES
ON THE ROBUSTNESS OF UNIDIMENSIONAL
NONLINEAR ITEM RESPONSE MODELS

DEGREE

MASTER OF EDUCATION

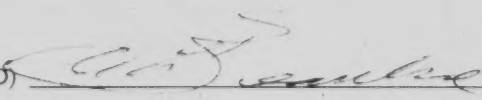
YEAR THIS DEGREE GRANTED

1990

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED)



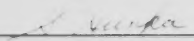
PERMANENT ADDRESS:

127 Greengrove Avenue,
Sherwood Park, Alberta
Canada T8A 3C5

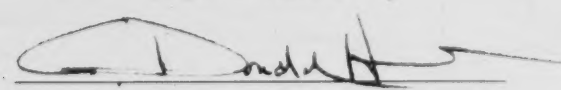
DATED 03 Aug. 1990

THE UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled THE EFFECT OF HIGHER ORDER LATENT SPACES ON THE ROBUSTNESS OF UNIDIMENSIONAL NONLINEAR ITEM RESPONSE MODELS submitted by ALBERT LEO BEAULNE in partial fulfilment of the requirements for the degree of MASTERS OF EDUCATION.



Dr. S. Hunka



Dr. D. Heth



Dr. J. McGregor

Date. 1990

Dedicated to
My Daughter, Janae Anne

Abstract

Latent Trait or Item Response Theory (IRT) relies heavily on a number of strong assumptions (Lord, 1980; Lord & Novick, 1968). Unidimensionality is considered to be the most essential of these assumptions (Hambleton, Swaminathan, Cook, Eignor and Gifford 1978). Several procedures now exist which estimate the parameters contained in the unidimensional IRT (UIRT) model. One which has received extensive use is the joint maximum likelihood procedure employed in the parameter estimation program Logist (Wingersky, Barton, & Lord, 1982).

The current study assessed the effects on the estimation of UIRT parameters when data sets violate the assumption of unidimensionality by exhibiting varying degrees of multidimensionality and correlations among the dimensions. The effects were assessed by generating data sets having two or three dimensions and correlations among the dimensions of 0.0, 0.3, 0.6, 0.95, and 0.99. Thus, ten data sets were generated, each representing a different combination of dimensionality and correlation among the dimensions.

Procedures for generating the data are described. The data were generated using Fortran 77 and IMSL subroutines. The suitability of the compensatory (CMIRT) and noncompensatory (NMIRT) multidimensional item response models used for generating the data are also noted. The problems encountered in generating the data and the techniques used to overcome those problems are described. Methods to ensure that the data sets did indeed contain the intended characteristics were of special interest.

The computer program Logist was employed to estimate the person and item parameters for the pseudo three parameter unidimensional IRT model (guessing parameter held constant at 0.2). The estimated parameters for each data set were compared to the parameters which were used in generating the data sets. The degree of congruity between the estimated parameters produced by Logist and the parameters inherent in the data sets was tested by examining the correlation between the IRT parameters, their means, and their

sums. Mean square differences were also examined to determine the size of the congruence when dimensionality and correlation were varied.

Increasing dimensionality had a negative impact on the congruence between the estimated parameters and the generated parameters. Conversely, increases in the correlation between the dimensions to some extent countered the negative effects of increased dimensionality.

Replication of Ansley and Forsythe's 1985 study, with respect to a two dimensional data structure, was completed and is presented as a part of the current study.

Acknowledgements

I wish to express my sincere appreciation for the assistance and close supervision I have received from Dr. Steve Hunka (thesis supervisor). I especially wish to thank him for the patient and noncritical way in which he guided me through the process of thesis preparation and completion. Further, I acknowledge the helpful comments and suggestions of Dr. D. Heth and Dr. J. McGregor

A special thank you is extended to the staff of the Division of Educational Research Services for the help and support they have provided over the last two years.

TABLE OF CONTENTS

CHAPTER 1.....	3
Introduction	3
Item Response Theory versus Classical Test Theory	3
Item and Person Parameter Invariance in IRT	6
Item Response Models.....	8
Assumptions	8
Unidimensionality.....	9
Local Independence	11
Item Characteristic Curves	12
The Logistic Model.....	12
Three Parameter Logistic Model: A Closer Look.....	15
Estimating Ability and Item Parameters.....	17
Conclusion	20
Purpose and Objective of the Study.....	21
CHAPTER 2.....	24
Review of the Related Literature.....	24
Factor Analytical Models	24
Summary	26
Item Response Models.....	26
Summary	27
CHAPTER 3.....	29
Research Design and Methodology	29
Model Selection.....	30
Compensatory Item Response Models.....	31
Noncompensatory Item Response Models.....	32
Parameter Selection	34

Data Generation.....	35
Simulation Model.....	37
CHAPTER 4.....	38
Analysis and Findings.....	38
Two Dimensional Data Structure.....	38
Data Confirmation.....	38
Data Analysis	40
Results.....	43
Ability.....	43
Discrimination	43
Difficulty	45
Three Dimensional Data Structure.....	46
Data Confirmation.....	46
Data Analysis	50
Results.....	51
Ability.....	51
Discrimination	53
Difficulty	53
CHAPTER 5.....	56
Summary, Conclusions, and Recommendations.....	56
REFERENCES.....	59

LIST OF TABLES

No.	Table	Page
1.	Descriptive Statistics for Two Dimensional Data	39
2.	Factor Analysis of Two Dimensional Data	39
3.	Eigenvalue Ratios for Two Dimensional Data	41
4.	Multidimensional Scaling for Two Dimensional Data.....	41
5.	Correlations and Absolute Mean Differences Between the Estimated and True Parameters for Two Dimensional Data	44
6.	Correlations and Absolute Mean Differences Between the Estimated and True Discrimination Parameters for Two Dimensional Data	44
7.	Correlations and Absolute Mean Differences Between the Estimated and True Difficulty Parameters for Two Dimensional Data	47
8.	Descriptive Statistics for Three Dimensional Data	47
9.	Factor Analysis of Three Dimensional Data	49
10.	Eigenvalue Ratios for the Three Dimensional Generated Data	49
11.	Multidimensional Scaling for Three Dimensional Data	49
12.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	52
13.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	52
14.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	54

LIST OF FIGURES

No.	Figure	Page
1.	Three Parameter Item Characteristic Curve.....	15

CHAPTER 1

Introduction

Item Response Theory versus Classical Test Theory

An item response model specifies a relationship between the examinee's observable test performance and the unobservable abilities (latent traits) assumed to underlie performance on the test. An example of such a model is the item-response function, called the item characteristic curve in the one-trait or unidimensional model, which relates the probability of getting a test item correct to the latent trait underlying performance on the items. Responses to a set of manifest variables (items), which are designed to assess performance on an achievement test, are less than perfectly related to the assumed latent ability. This suggests the presence of one or more underlying mental traits. In the case of continuous manifest variables the problem of searching for latent traits leads directly to factor analysis.

The relationship between IRT and factor analysis is well established (see Traub & Wolfe, 1981), and the unidimensional latent trait in IRT presumes a one-factor structure. A one-factor analytic model is sufficient for the two-parameter normal ogive IRT model (Lord & Novick, 1968). Factor analytic procedures are useful for testing a priori structures and for determining the appropriate IRT models (e.g., Hulin, Drasgow, & Parson, 1982). IRT approaches the problem of determining the relationship between a response and an ability from a probabilities viewpoint rather than from a correlational one. That is, the function that specifies the relationship between an item and the latent trait is stated in terms of cumulative probabilities for given trait values.

Commonly, IRT models are associated with multiple-choice items used in an ability or achievement test (Traub & Wolfe, 1981; Warm, 1978). Most multiple-choice test items form dichotomous scales whose categories can be labeled "0" for incorrect responses and "1" for correct responses. Many other item types, including some in which subjects supply or construct the response, are also dichotomous; any item type where the subject's

response is marked correct or incorrect or that can be scaled as two points is dichotomous. Items that can be scored in this way are quite common, and the IRT model therefore applies rather broadly. Many of the recent textbooks on IRT deal almost exclusively with IRT models based on dichotomous types of test items.

IRT models are not limited to multiple-choice items (or to dichotomous manifest variables generally); work is being done with multidimensional item types as well (Mislevy, 1987, p. 240). For example, researchers have applied IRT models to nominal items with more than two categories and to items with ordered scales: e.g., Andrich (1978), Bock (1972), Masters and Wright (1984), and Samejima (1972). Others have extended the notion of IRT to multidimensional models (Bock & Atkin, 1981; Doody-Bogan & Yen, 1983; Hattie, 1981; Mulaik, 1972; Rasch, 1961; Reckase, 1985; Samejima, 1974; Simpson, 1978; Whitely, 1980).

IRT can provide to test producers and test users certain benefits which are not found in Classical Test Theory. The statistical indices found in Classical Test Theory are typically those deriving from norm-referenced techniques (see Baker, 1977, for a review of norm-referenced item analysis statistics) and are defined below.

- 1). Item difficulty--usually indicated by the "p-value" which is the proportion of examinees with a correct response to the item (proportion scoring "1").
- 2). Item discrimination--usually indicated by the point-biserial or biserial correlation of the item with the total test.
- 3). Average and spread--usually mean and standard deviation, but sometimes also median, semi-interquartile range, etc.
- 4). Distribution--usually skewness and kurtosis of the distribution, but goodness of fit to the expected distribution can also be tested.
- 5). Reliability of the test scores--usually KR-20 or corrected split-half (which under estimate reliability based on the classical test theory notion of strictly parallel tests), or, often in ability testing, the test-retest correlation.

- 6). Error associated with a score--usually the standard error of measurement as calculated from the standard deviation and the reliability estimate.

Hambleton & Swaminathan (1985) identify five major shortcomings of the classical test theory approach to test development and test evaluation (see also Hambleton & van der Linden, 1982). The first is related to the examinee sample on which the statistics are calculated. The item p-value is directly affected by the ability level of examinees, and it is not necessarily equal for the same item administered to two groups of the same average ability. The test characteristics such as average, spread, and form are dependent on the ability level of examinees as well. Item discrimination is related to subject matter homogeneity, range of examinee ability scores, and the dispersion of p-values. The discrimination of items affects the test reliability as measured by internal consistency indices (e.g., KR-20) and hence the typically reported standard error of measurement.

The second shortcoming of using classical test theory is related to making comparisons among groups, specifically that the same test or parallel tests must be used (Hambleton & Swaminathan, 1985). The test cannot be adapted to the examinee, and for high or low ability subjects the test is usually less precise. Further, test validity can be increased by matching item difficulty to the ability of the examinee (Lord, 1980). Comparison of scores from two or more different tests becomes very difficult under the classical model, even for group averages relatively close to the middle of the score scale. The problem of precision of scores being different at various points of the score scale is masked in classical theory by the common practice of employing one estimate of the standard error of measurement for all levels of the test scale. Hambleton and Swaminathan (1985) posit that performance at the high end of the score scale is frequently more stable than at the middle and lower ends.

The third shortcoming expressed by Hambleton and Swaminathan (1985) is that test reliability is defined in terms of parallel forms. Parallel tests are most often difficult to

achieve, and most reliability indices are either lower-bound estimates (e.g., KR-20) or with unknown bias (e.g. test-retest).

A fourth shortcoming is that classical test theory cannot provide a basis for predicting examinee performance on a given test item, that is the probability of responding correctly cannot be estimated from item statistics. The ability to predict item performance is necessary for tailored testing and is desirable in many testing situations.

Classical test theory has fallen short with respect to identifying biased items and in equating test scores (Hambleton and Swaminathan, 1985, p. 3). This follows from the criticisms specified above, but it indicates problem areas for which solutions are clearly aided by IRT. Hulin, Drasgow, and Parsons (1982, p. 8) present the argument that the total test score is unrepresentative of ability. Patterns of response are lost in this practice, and differential weighting of items and/or individually selected items (adaptive testing) may provide a more accurate score. Clearly, both differential item weighting and individual item selection based on estimates of examinee ability and item parameters can produce more accurate examinee scores, and this accuracy can be obtained more efficiently (by fewer items). This fact is readily apparent from combining item information functions and comparing test information at various points on the ability continuum. Certain combinations of items produce much more information than do other combinations at specific points on the scale (Warm, 1978, pp. 73-77, provides an example of this from actual IRT item data).

Item and Person Parameter Invariance in IRT

IRT is a theory which relates the observed performances of examinees on items to a latent trait, which in the unidimensional case, is said to explain the behaviour on the items. The relationship is described as a probability function; this function depends on the latent ability of a person, a single parameter in the unidimensional case, and on information about the item, one, two or three parameters, depending on the model (Traub & Wolfe, 1981). However, in contrast to Classical Test Theory in which total test score is used to estimate

the latent ability, IRT uses all the parameter information from every item administered to the examinee in determining the level of ability.

In contrast to the shortcomings mentioned in relation to Classical Test Theory, IRT is claimed to provide item and test statistics which are independent of examinee characteristics and of test characteristics (Warm, 1978, p. 17; Lord & Novick, 1968). As well, the fit of the item models and test models can be evaluated empirically. The following features of item response models are outlined by Hambleton and Swaminathan (1985, p. 11) and are applicable when the IRT model fits the data. First, item parameter estimates are independent of the group of examinees sampled from the population of examinees for whom the test was designed. Second, examinee ability estimates are independent of the particular choice of test items sampled from the population of items which were calibrated (the calculated item parameters). Third, the precision of ability estimates is known.

It should be noted that some authors are not confident about the "robustness" of IRT models, and emphasize the problems associated with the strong assumptions necessary for their application, in particular unidimensionality and local independence (Goldstein, 1980; Traub, 1983; Traub & Wolfe, 1981). Bock, Mislevy, and Woodson (1982) recommend IRT functions based on logical learning units rather than combining a variety of skills and employing one dimension to define the model, for example, when a test of mathematics contains items from a variety of objectives and is analyzed using one dimension.

Traub and Wolfe (1981) comment that if an IRT model is correct it provides tremendous advantages, but in the practice of measuring educational achievement the assumptions are rarely tenable. They also point out that present-day IRT models require dichotomous items and assume a single underlying dimension across a range of skills or even grade levels. They are particularly concerned that IRT models are not independent of the context: "we view as potentially dangerous the practice of applying latent trait scaling over time and over educational programs where instruction varies" (Traub & Wolfe, 1981, p. 380). Traub (1983) provides data that demonstrate the unlikelihood of the

unidimensionality assumption. Traub & Wolfe (1981) point to a further problem, which is that the test of adequacy of fit of the model has statistical deficiencies, such as lack of power, nonlinearity in the data, and poor generalizability of item parameters.

Because of the difficulty and expense associated with using IRT in assessment, most present applications tend to be in large-scale achievement and ability testing programs, either government-based achievement testing programs such as the California Assessment Program (Pandey & Carlson, 1983) or in published standardized ability and achievement tests such as the Comprehensive Tests of Basic Skills by McGraw-Hill (Yen, 1983). Large-scale testing programs almost exclusively utilize multiple-choice test items. Thus, applications of IRT have been made with this type of item. However, it is well known that multiple-choice items cannot assess much of what is important in education (e.g., writing competency), and therefore the application of IRT, as it has been developed, is very limited and may even be limiting to good assessment practices.

Item Response Models

A number of models have been proposed that specify the expected relationship of the latent trait to the categories of the items (the manifest variables). The goodness of fit of each can be established and comparisons can be made (Hambleton & Swaminathan, 1985). Traub and Wolfe (1981) express the concern, "the assessment of model fit is something that, in practice, is usually done very badly" (p. 384), and the power of statistical tests is usually poor. The models are based on various assumptions about the data, and a comparison of the fit of two models becomes a test of the particular assumptions that distinguish the two models.

Assumptions

There are three assumptions that apply generally to latent-trait models, and more specifically to IRT and the item response function. The first is related to the dimensionality of the latent space; it is assumed that only one ability or trait is necessary to explain an examinee's test performance. The second is related to local independence, i. e., examinees

responses to different items in a test are statistically independent at a given value of the latent trait. The third is related to the item characteristic curves; the curve connecting the means of the conditional distributions is the regression of item score on ability (referred to as an item characteristic curve).

Unidimensionality. Item response models which assume a single underlying trait are referred to as being unidimensional. In a strict sense this assumption is only met in theory and is never realized in practice. In any testing situation a host of cognitive and motor skills all come into play and influence the outcome of any one subject's performance on a test. What is considered to be of importance in meeting the assumption of unidimensionality is that a single "dominant" ability factor underlies performance on a test and that the test is designed to measure the ability in question (Hambleton & Swaminathan, 1985). Warm (1978), presents three rules of thumb for determining unidimensionality.

1. Tests that look unidimensional probably are unidimensional.
2. Items that test bits of knowledge that were learned together are probably unidimensional.
3. Items that test bits of knowledge which are logically and sequentially related are probably unidimensional.

Warm cautions that such rules of thumb are meant to be nothing more than a guide and are not presented as replacements for sound empirical evidence for unidimensionality. Warm posits that such rules of thumb are justifiable given the difficulty in determining unidimensionality empirically. The difficulty arises from the fact that most tests of unidimensionality rely on factor analysis of inter-item tetrachoric correlations, which requires that the ability levels be normally distributed. This assumption need not hold when items allow some probability of a correct response through guessing (Warm, 1978).

A unidimensional instrument is often considered in terms of the outcome of a procedure designed to establish its singularity. However, a unidimensional test is not defined in relation to a unit rank, deviations from a perfect scale, or one common factor.

Unidimensional tests are not necessarily reliable, internally consistent, or homogeneous (Hattie, 1985). The only defining character is that performance on a test be a function of a single underlying ability.

Critics of unidimensional test construction have pointed to a number of serious flaws in the attempt to develop homogeneous test (Humphreys, cited in Harrison, 1986; Traub, 1983). Harrison is concerned that tests which are constructed to be unidimensional may be too limited in their capability to assess the subject in a given area. He further stresses a paradoxical situation in which as a test becomes more homogeneous with respect to a given underlying trait it also decreases in predictive validity.

Traub (1983) refers to the notion of a singular underlying trait as "the fragile assumption of unidimensionality". He considers three circumstances which could affect the dimensionality of the latent ability space. The first is the method of instruction. Differential instruction can create a multidimensional ability space where before a unidimensional ability space existed. The simple example presented by Traub is that of children who are homogeneous with respect to the elementary operations of adding and subtracting integer numbers. The children are then split into two groups; the first group receives exclusive instruction in addition while the second receives exclusive instruction in subtraction. When the two groups are later tested on both operations there is a negative correlation between pairs of items relating to addition and subtraction. This would imply two abilities rather than a single ability which existed prior to the differential training.

The second circumstance is that of a speeded test. As before there is a single group of children with the same ability to answer the questions on the test. Again they are split into two groups, such that they receive differential instruction in test taking or such that they are divided on some criterion which allows them to be grouped as fast or slow test takers. As individual groups their inter-item correlations are zero but, when combined the correlation is greater than zero. A non-zero correlation, for a subpopulation of examinees

all of equal ability on one latent trait, implies that a single trait is not sufficient to satisfy the condition of local independence.

The scenario is virtually identical for the third circumstance, which is guessing. If individuals of equal ability can be divided into two groups, one with a propensity to guess and the other with the absence of guessing, then, as before, the inter-item correlations within the groups will be zero. However, there will be an inter-item correlation for the combined groups. The same conclusion can be drawn as for the case of a speeded test, that is, more than one underlying trait is responsible for performance on the test.

Traub concludes that no unidimensional model can adequately represent achievement data. He further states that it would be foolhardy to expect errors due to model misfit to average to zero as random errors would. Therefore, he suggests that one should seek other solutions to educational measurement problems.

Local Independence. There is a clear relationship between local independence and the concept of unidimensionality (Bejar, 1980; Hambleton & Swaminathan, 1985; Traub & Wolfe, 1981). Local independence states that the joint probability of the scores on two items for a given ability level is the product of the probabilities of the score on each item given the ability level (Traub & Wolfe, 1981, p. 386):

$$\Pr(X_i=x_i \text{ and } X_j=x_j | \theta) = \Pr(X_i=x_i | \theta) \times \Pr(X_j=x_j | \theta)$$

where X_i and X_j are scores on any two items and θ is any given ability level. This rule can be simply extended to the joint probability of a response pattern on any number of items. The equation implies that any correspondence between pairs of items must have been accounted for by the latent trait, but it may be that several traits are needed for this condition to be satisfied. In this sense, then, a unidimensional test must have local independence of item responses. However, the inverse is not true, that is local independence does not imply unidimensionality. Conditions which may violate the assumption of local independence are those of a speeded test (discussed above) and chained items.

Presumably, factor analysis of item responses can be used to test the reasonableness of this

assumption, and Lumsden (1976) suggests the ratio of first- to second-factor variance as an index. This procedure has its problems since it relies on item correlations (see Hambleton & Swaminathan, 1985, pp. 21-22; Warm 1978, pp 99-101).

Item Characteristic Curves. The model is typically expressed in terms of the *item characteristic curve* (ICC), or the item characteristic function in the multidimensional case, which is the (nonlinear) regression of the item score on the latent trait or ability. Since the probability of an examinee answering an item correctly is dependent only on the form of the ICC, it is independent of the performance of other examinees, and therefore the curve is invariant across samples (Bejar, 1980; Hambleton & Swaminathan, 1985). The ICC is the curve relating the probability of item i being correctly answered by a randomly selected individual with ability θ . Alternative definitions are possible but may lead to problems; this is the suggested definition of Hambleton & Swaminathan (1985, p. 27). The ability scale θ is not defined by the item scales, but is usually arbitrarily set to have mean = 0 and standard deviation = 1, thereby making the practical range approximately from -3 to +3. Finally, the distribution of θ does not need to be normal for the theory (Warm, 1978).

The Logistic Model

A wide variety of models is possible to define the item characteristic function $P_i(\theta)$ for typical achievement or ability tests but the model commonly in use today is the logistic one, attributable to Birnbaum (1968). This model is preferred over the normal ogive originally proposed by Lord, since it is more mathematically tractable, i.e., it does not involve integration in the equation for $P_i(\theta)$. The two-parameter logistic model, with the scaling factor $D = 1.7$ (see formula below), can be shown to produce an item characteristic curve which deviates from that produced by the normal ogive by less than .01 for all values of θ (Hambleton & Swaminathan, 1985, p. 37). The logistic model ICC for item i favoured by many writers today (e.g., Warm, 1978) involves the following three parameters:

a_i = the *item discrimination* value of the ICC: it is proportional to the slope of the ICC at the point on the curve where $\theta = b_i$; $-\infty \leq a \leq +\infty$

b_i = the *item difficulty* value or location of the ICC: the point on the θ axis where the examinee has a $(1+c_i)/2$ chance of responding correctly to the item (also at the inflection point of the curve); $-\infty \leq b \leq +\infty$

c_i = the *pseudo-chance level* of the ICC: the lower asymptote value of $P_i(\theta)$, or the "guessing factor" in multiple-choice items. $0 \leq c \leq 1$

The three-parameter equation for $P_i(\theta)$ can be expressed as follows:

$$P_i(\theta) = c_j + \frac{(1-c_j) e^{Da_j(\theta-b_j)}}{1 + e^{Da_j(\theta-b_j)}}$$

where $D = 1.7$ (the scaling factor), and $i = 1, 2, \dots, n$ items. This may also be expressed as:

$$P_i(\theta) = c_j + (1-c_j) [e^{-Da_j(\theta-b_j)}]^{-1}$$

A four-parameter model has also been posited, the fourth parameter being an upper asymptote less than one for $P_i(\theta)$ based on the notion that even high ability examinees miss items through carelessness (see Hambleton & Swaminathan, 1985, pp. 48-49; also Traub & Wolfe, 1981, p. 423). The three-parameter model is the one most commonly discussed, and Lord provides some evidence that there is little practical gain from adding the fourth parameter (Hambleton & Swaminathan, 1985, p. 49).

Based on this general IRT model there are two restricted models, namely that of one parameter, b_i , and that of two parameters, a_i and b_i . There is considerable debate in the literature regarding the appropriate model to use in actual applications. It seems obvious that the three-parameter model should be most applicable to multiple-choice items since it

includes the possibility of both varying discrimination of items (a_i) and the guessing factor (c_i). However, there is considerable use of the one-parameter model; this is equivalent to the Rasch model which assumes that the discrimination parameter is constant for all items (see Wright, 1977). The Rasch model is easier to use since it involves estimation of only one parameter for each test item, and numerous applications are reported in the literature (see Wright & Stone, 1979).

In addition to its use with multiple choice items, the two-parameter model is also appropriate for test items which are of a constructed-response format, scored right or wrong, and where there is little or no opportunity for the examinee to respond correctly by guessing. Some of the original work by Lord (e.g., 1953) was on the two-parameter normal ogive model, and Birnbaum (e.g., 1968) proposed the two-parameter logistic model as an alternative to the normal ogive.

A variety of IRT models have been developed, but the one, two, and three-parameter logistic models are most commonly used. These models can be applied to the item responses and the fit of each can be compared to that of the others. The scale of θ is arbitrarily established and can be adjusted to a convenient metric, such as mean = 50 and sd = 10 (T-scale), by using a linear transformation. One such transformation for the three-parameter ICC is:

$$\theta^* = g\theta + k, \quad c_i^* = c_i, \quad b_i^* = gb_i + k, \quad \text{and} \quad a_i^* = a_i/g,$$

where the scale size is adjusted by g , a location shift of k is made, and the constant c_i^* is arbitrarily set for the item. Hambleton and Swaminathan (1985) demonstrate that this form of transformation leaves $P_i(\theta)$ invariant, i. e., $P_i(\theta^*; a^*, b^*, c^*) = P_i(\theta; a, b, c)$. These transformations can be applied to the two-parameter model, and to the one-parameter model if a_i is taken to be the average item discrimination, say a . If a computer program returns θ in terms of mean = 0 and sd = 1, as many programs can do (e.g., LOGIST; see Wingersky, 1983), then to transform to a T-scale simply means $g = 10$ and $k = 50$.

Three Parameter Logistic Model: A Closer Look

The three-parameter logistic model can be used to develop some interesting relations and applications of IRT. The ICC for item i is:

$$P_i(\theta) = c_i + (1 - c_i) \left[e^{-Da_i(\theta - b_i)} \right]^{-1}$$

From the equation it is readily apparent that the value of $P_i(\theta)$ approaches c_i as θ approaches $-\infty$ and 1 as θ increases to $+\infty$. For example, when $\theta \rightarrow +\infty$, the expression $[-Da_i(\theta - b_i)] \rightarrow -\infty$ and the limit of e raised to the power $-\infty$ becomes 0, so that $1 + \exp[-Da_i(\theta - b_i)] = 1$ and $c_i + (1 - c_i)/1 = 1$. The probability of an examinee with infinite ability responding correctly to the item is 1, as expected.

The relationship of $P_i(\theta)$ to θ resembles a cumulative distribution function (ogive) for examinees of all ability levels.

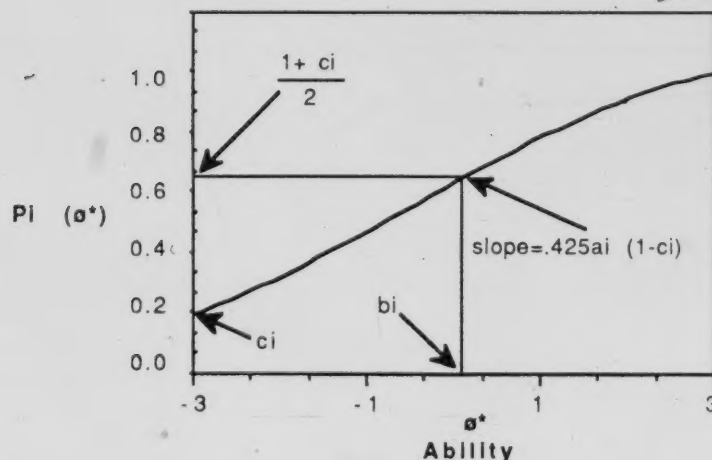


Figure 1. Three Parameter Item Characteristic Curve

Figure 1 makes some of the significant features of an ICC apparent. The value for c_i gives the minimum or lower asymptote for $P_i(\theta)$, whereas the upper limit is 1. The ICC increases monotonically from $P_i(-\infty) = c_i$ to $P_i(\infty) = 1$, with the greatest rate of increase occurring midway between these two extremes. Since the shape of the curve of the ICC is

centro-symmetric about the point $P(b_i)$, b_i is the value for θ at which the inflection point of the curve occurs. This is the "midpoint" of the curve, and thus $P(\theta=b_i) = (1 + c_i) / 2$. This point is also the point of maximum slope of the curve, and thus the maximum value of the discrimination, a_i , for item i . This implies that item i discriminates most among examinees of differing abilities where these abilities lie close to $\theta = b_i$, i. e., at $\theta \approx +0.5$ in Figure 1 above (when $c_i = 0.0$).

The foregoing are two crucial aspects of IRT. First, the slope of the ICC for an item is limited by the value for c_i , with the maximum obtaining when $c_i = 0$. As Warm (1978) puts it, "the effect of the c -value is to squeeze the ogive into a smaller vertical range. . . . equal to $1 - c$ " (p. 32). This can be visually represented by graphing an example where c_i approaches 1.0; the ICC would be quite flat and the slope $\rightarrow 0$. This implies that items which are prone to considerable guessing, such as true-false items, would likely be less discriminating than those where guessing is at a minimum, such as constructed-response items. Second, an item's capability to distinguish between examinees at adjacent ability levels is dependent on what point along the ability continuum is being considered, with the greatest discrimination occurring at the point b_i . This implies that certain items would be more discriminating at particular levels of ability than others. For example, given that the parameters a and c were equal for two items, the item with a value for b closest to the ability levels for which discrimination was desired would be the most discriminating. The item whose ICC is depicted in Figure 1 above is indicative of an item which has very little discriminating power.

These features of IRT are in striking contrast to those of classical test theory, where the effect of guessing has little impact on the item discrimination, unless it becomes rather large, and item discrimination is assumed to be constant irrespective of examinee ability. The impact of the difference becomes even more significant when a test consisting of a collection of items is considered. The application of IRT to tailored testing is obvious: it allows selection of items based on their power to discriminate most at the estimated level of

ability of the examinee. Since in many cases there is no estimate of examinee ability it can simply be inferred from performance on previous items, and the testing can be started by administering moderately difficult items.

A few comparisons of the parameters to classical theory statistics are given by Warm (1978, pp. 51-53). Unfortunately, the relationships are complex, so will not be described here. What can be given is some indication of commonly obtained values for the parameters. Warm (1978) contends that items with a -values below .80 are not sufficiently discriminating for most purposes. He also states that b -values below -2.5 are quite easy and above +2.5 are very difficult. Finally, he argues that a reasonable estimate of c for a multiple-choice item is given by $(1/A) - .05$ where A is the number of options, and that although c 's typically range from 0.0 to 0.4 they should be 0.2 or less. A *test characteristic curve* can also be determined simply by the averaging the $P(\theta)$'s for all items in the test (see Hambleton & Swaminathan, 1985, pp. 61-69). This produces a curve that is similar to the item ICC's. It is further defined as the relationship of test true score (or domain score) to θ .

Estimating Ability and Item Parameters

Ability (θ) and item parameters (a , b , and c) must be simultaneously estimated from the response patterns of the examinees. If N examinees respond to n items, the number of parameters to be estimated are $N + 3n$, and $N + 2n$ and $N + n$ parameters for the three, two and one parameter models respectively. There are indeterminacies in the solutions so restrictions must be applied. Usually for θ in the three and two parameter models the mean and standard deviation are set to 0 and 1 respectively, which reduces the parameters to be estimated by two.

The algorithm for the maximum likelihood solution for the IRT parameter estimation problem is beyond this paper. It is described briefly in Warm (1978), and in somewhat greater detail in Hambleton and Swaminathan (1985) and Lord (1980). The procedure involves maximizing the likelihood function which is based on the conditional probability

of the particular response vector $\mathbf{u}$ for examinees on the items, given the ability parameter and item parameters, i. e.,

$$L(\mathbf{u} | \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c})$$

where the terms are as defined in the next paragraph. The function is obtained by taking the product of the joint probabilities of the responses $\mathbf{u}$ and the parameters (this is possible because of the local independence assumption). But since usually all of the parameters are unknown (actually all but two), they must be estimated simultaneously by an iterative procedure.

The natural logarithm of the likelihood function (Hambleton & Swaminathan, 1985) is obtained and since the function is a product of probabilities, the result becomes a summation:

$$\ln L(\mathbf{u} | \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_j \sum_i [u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln Q_{ij}]$$

where:

Summation is over N examinees ($j=1, \dots, N$) and n items ($i=1, \dots, n$)

$\mathbf{u}$ is an Nn vector of the N examinee observed responses to the n items (element u_{ij} is the score of examinee j on item i)

$\boldsymbol{\theta}$ is the vector of N ability estimates

$\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the three vectors of parameters for the n items

$P_{ij} = P_i(\theta_j)$ is the ICC value for item i and ability level θ_j

$$Q_{ij} = 1 - P_{ij}$$

The maximum likelihood equations are obtained from the partial derivatives of L with respect to each parameter vector (a vector consists of one set of parameters $\boldsymbol{\theta}$, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$). Solutions are obtained for the equations by first treating the item parameters as known and producing N equations in θ_j . An iterative procedure is employed until differences between successive approximations to the item parameters becomes sufficiently small.

Solutions for tests with substantial numbers of items administered to large samples are a difficult problem given the large number of nonlinear equations which are involved,

particularly in the case of two and three parameter models. Several additional problems arise beyond the sheer difficulty of computing. The first is that the numerical procedures do not guarantee an absolute maximum for nonlinear equations and local maxima may be obtained (Hulin, Lissak, & Drasgow, 1983). Second, estimates of the parameters may take on values outside the acceptable range, and researchers such as Wright (1977) argue that this calls into question the maximum likelihood solution procedures for two and three parameter models. Third, the estimate for the lower asymptote (c) is very difficult to achieve using maximum likelihood methods (Lord, 1980).

The computer program LOGIST (Wingersky, 1983) employs this joint maximum likelihood estimation procedure. Hambleton and Swaminathan (1985) and Mislevy (1987) state that it is one of the computer programs commonly used for the two and three parameter problem, but indicate that there is interest in alternate procedures such as those based on marginal probability functions. A computer program using this approach is BILOG. The authors also describe conditional maximum likelihood estimation, which with the proper constraints leads to the Rasch procedure of Wright and Stone (1979) and to the program BICAL. Although a variety of other solution procedures are being considered, one that appears quite promising is Bayesian estimation (Hambleton & Swaminathan, 1985). Apparently, this approach produces modest improvements on the point estimates (Mislevy, 1987). Lord (1980) suggested samples of more than 1000 examinees and more than 50 items for adequate parameter estimation. This seems to be good advice particularly where ability estimates are to be used for decision-making: Mislevy (1987) states "that treating estimates as parameters provides fairly accurate end results in applications when both n and N are large--say, $n > 50$ and $N > 2000$ " (p. 253). He also states that serious biases result from treating θ estimates as parameters or true scores when examinees take few items: "less than, say, 15" (Mislevy 1987, p. 255) and few items is precisely what tailored testing would hope to achieve! The problem of obtaining estimates of θ given item parameters is much simpler than that of simultaneously estimating examinee and item

parameters. Hambleton and Swaminathan (1985) describe the procedure, and Warm (1978) gives a simple example of how to obtain estimates of θ for three items.

Conclusion

Consideration has been given to many of the shortcomings of classical test theory: a) the variability of item statistics as a function of the sample group, whereas, in IRT item parameters remain invariant across groups of examinees and the ability parameters are not affected by the items administered, b) the difficulty in comparing groups tested on items of different difficulty levels, and the problem of achieving parallel measures, c) the lack of predictive power when trying to ascertain the probability of a subject's response to a particular item, d) the assumption of homogeneity of error variance among subjects who are heterogeneous with respect to ability. In addition to these, classical test theory has been unable to provide satisfactory information with respect to identifying biased items, equating of test scores, and development of tests which discriminate maximally at a given ability level.

Item Response Theory addresses many of the problems associated with Classical Test Theory. The invariance of both item and person parameters makes it possible to equate individual scores on tests of differing difficulty levels. Either the ability scores or transformed scores may be reported. Item response models allow for the detection of item bias through an inspection of the item response function for the groups in question and utilization of significance tests. The construction of a highly discriminating test can be accomplished through an inspection of the item parameters. For example if a test is needed to distinguish between subjects at the high end of the ability range, then items with large value of "b" and "a" would be selected.

Item Response Theory is not without its criticism. The assumption of unidimensionality in particular has been attacked on several levels. It would appear that a

move to multidimensional models will be a natural transition in order to eliminate some of the stronger arguments.

Purpose and Objective of the Study

The process of developing mathematical models to describe and understand empirical phenomena is common in many areas of natural science. Einstein's mathematical model relating energy and mass is just one of the many examples. Mathematical models have been developed to study the flow of substances, from the blood through the human body to the flow of traffic in our inner cities. Mathematical models have also been employed to understand the nature of hurricanes and earthquakes, so that one may predict their occurrences and better be able to minimize their negative consequences on the environment. They have also been employed extensively in the study of population growth. For example the Volterra Model is a nonlinear model of interacting populations. When the data fit the model these mathematical functions become powerful tools which can predict the behaviour of subatomic particles, as in the case of models related to physics, or save the lives of people, as in models related to medicine.

Psychology has made extensive use of mathematical models to describe the behavior of organisms. Some of the earliest mathematical models were developed by Ernst Weber in 1834 and Gustav Fechner in 1860. Herstein (1961) developed a model to describe the matching behavior of animals. Baum (1974) extended the Herstein model to encompass deviations from the matching law. Rescorla and Wagner (1972) as well as Pearce and Hall (1980) presented mathematical models to describe the process of learning. Iwasa, Higashi, and Yamamura, (1981) presented a model to describe how animals exploit food distribution in a patch. Tatsuoka (1968) gives a more complete discussion of the development and use of mathematical models in psychology.

Item response models employ a mathematical function to describe the relationship between an observable response and an unobservable ability. Employing a mathematical function to describe this relationship classifies IRT models as mathematical models (Hambleton & Swaminathan, 1985; Tatsuoka, 1968). At some stage in the development of a mathematical model certain simplifying assumptions are made. These assumptions usually increase the model's mathematical tractability. With respect to unidimensional item response models, unidimensionality is one such simplifying assumption. A definition of unidimensionality has been somewhat controversial (Hambleton & Rovinelli, 1986). A more complete discussion will be given in the section on unidimensionality. It should be noted that this same assumption is also an underlying tenant of classical test theory. The importance of this assumption to current IRT models has stimulated a large number of attempts to develop indices of unidimensionality. Hattie (1985) has summarized 87 such indices. The most promising determinate of unidimensionality appears to come from analysis of the absolute sums of squares of residuals and the number of residuals greater than some criterion value resulting from nonlinear factor analytical techniques.

Given the difficulty of developing adequate indices of unidimensionality, a logical step would be to assess the robustness of unidimensional models when a data set has a multidimensional structure. To date, several studies have assessed unidimensional models under just such a condition. Some have approached it from a factor analytical point of view (Drasgow & Parson, 1983; Harrison, 1986), while others have approach it from an IRT framework (Ackerman, 1989; Ansley and Forsyth, 1985; Reckase, 1979; Way, Ansley, & Forsyth, 1988).

The current study will extend the research aimed at assessing the issue of the robustness of UIRT models to violations of the assumption of unidimensionality from within an IRT framework. However, where the previous research focused on two

dimensional data structures, the present study will attempt to ascertain whether the findings of these former studies can be generalized to data sets generated to have a three dimensional ability space. These are data sets in which the probability of a correct response to any one item for a given individual is a function of that individual's current level on three different ability (θ) scales. The importance of this is to ensure that the results obtained, by employing a two dimensional data structure, are not an artifact of some unique relationship which exists within a two dimensional system. Further, it is likely that human behaviour is more complex than is suggested by a one or two dimensional model. It is therefore necessary to explore whether these results will hold when extended to more complex latent spaces.

CHAPTER 2

Review of the Related Literature

Item response models which assume a single underlying trait are referred to as being unidimensional. In a strict sense this assumption is only met in theory and is not realized in practice (Ackerman, 1989; Ansley & Forsyth, 1985; Birenbaum & Tatsuoka, 1982; Drasgow & Parson, 1983; Harrison, 1986; Reckase, 1979; Way, Ansley, & Forsyth 1988). In any testing situation a host of cognitive and motor skills all come in to play and influence the outcome of any one subject's performance on a test. What is considered to be of importance in meeting the assumption of unidimensionality is that a single "dominant" factor underlie performance on a test and that the test is designed to measure the ability in question (Hambleton & Swaminathan, 1985).

Recently, a number of studies have been conducted to assess the robustness of IRT models to the violation of the assumption of unidimensionality of the latent space. The general approach has been to simulate multidimensional data sets via a mathematical model. For some a linear factor analytical approach has been utilized in order to produce a response matrix (Drasgow & Parson, 1983; Harrison, 1986; Reckase, 1979). Others have employed multidimensional extensions of existing unidimensional IRT models (Ackerman, 1989; Ansley & Forsyth, 1985; Way, Ansley, & Forsyth 1988).

Factor Analytical Models

Drasgow and Parson (1983) simulated correlated common factors through a hierarchical factor analysis model developed by Schmid and Leiman (1957). The authors concluded that if the dominant latent trait is sufficiently strong then unidimensional IRT models provide an adequate representation of multidimensional data. However, if a single dominant latent trait is not sufficiently potent, then the use of a unidimensional model is inappropriate. Specifically, their results indicated that the general latent trait could be

successfully recovered if the correlation between common factors was greater than or equal to 0.50.

Harrison (1986) followed many of the the same procedures outlined in the Drasgow and Parson (1983) study. However, Harrison (1986) conducted his study employing a variety of patterns of factor loadings for the common factors, the test length, and the correlation between the common factors. The results reported by the author were compatible with those presented by Drasgow and Parson (1983). As the strength of the second order general factor increased the Logist parameter estimates improved relative to the theoretical parameter values. This same positive effect on the parameter estimate was seen as test length increased and for items uniformly distributed within the common factors versus items displaying a skewed distribution. Harrison concluded that as a single group factor controls variation in more items and in a larger number of items then the Logist program takes this factor as part of the unidimensional trait. Further, Harrison posited that the Logist estimation procedures were robust to violations of the assumption of unidimensionality, even when common factors display only moderate intercorrelations and that Logist is able to successfully recover parameters implied by second order general factors.

Reckase (1979) investigated the effect of multidimensional data on the one and three parameter logist model. Five simulated and five real data sets were employed in the study. When the data set contains more than one independent factor the 3 parameter model discriminates among ability levels on one factor and ignores the rest. The 1 parameter model estimates represent the sum of the factors. When the data set contains one large factor with a number of smaller factors both models measure the first factor, and the size of the first factor affects the parameter estimation in a positive way. Reckase posited that the one parameter and the three parameter models measure different abilities when independent factors are inherent in the data, but both measure the first principal component when the

first factor is large relative to the other factors present and this first factor should account for at least 20 percent of the test variance in order for parameter estimates to be stable.

Summary

The results from the above mentioned studies tend to support the hypothesis that unidimensional IRT models do not require that the latent ability space exhibit a singular structure in order to provide valuable information about the nature of the item and trait parameters presented in the model. However, it is also clear that a single dominant factor must underlie the test data. A number of problems present themselves when a factor analytical approach is employed to represent the structure inherent in the response matrix. The tacit assumption is that the responses have a linear relationship to the underlying trait. This is contrary to the nature of IRT models (Hattie, 1985). A further assumption is that multidimensionality is expressed across a set of items as opposed to within the items as an IRT model might suggest. The following three studies address these concerns by employing multidimensional IRT models to simulate the item responses.

Item Response Models

Ansley and Forsyth (1985) simulated multidimensional test data (two dimensions) using a noncompensatory multidimensional item response model (NMIRT). The model was first presented by Simpson (1978) as an extension to the three parameter unidimensional logistic model (Birnbaum, 1968). Ansley and Forsyth reported that the Logist estimates for the discrimination parameters a^* and the ability parameters ϕ^* were best represented as the average of the true a_1, a_2 and ϕ_1, ϕ_2 values respectively while the estimated difficulty parameters b^* were an over estimate of the b_1 value.

Way, Ansley, and Forsyth (1988) compared the results of Logist parameter estimates when the data had been simulated using both a compensatory (Doody-Bogan & Yen, 1983) and a noncompensatory (Simpson, 1978) model. The Doody-Bogan and Yen model is, as is the Simpson model, an extension of the three parameter logistic model (Birnbaum, 1968). The result of using the noncompensatory model to simulate data was comparable to

those reported by Ansley and Forsyth (1985). For the compensatory model the estimated difficulty parameters b^* and the ability parameters θ^* were best represented as the average of the true b_1, b_2 and θ_1, θ_2 values respectively while the discrimination parameters a^* was best represented as a sum of the true a_1, a_2 values. In all cases as the correlation between the dimensions increased the latent space became more unidimensional.

Ackerman (1989) in a simulation study, somewhat the same as that conducted by Way et al. (1988), compared the results of Logist parameter estimation when the data had been simulated using both a compensatory and noncompensatory models. Ackerman considered the additional effect on parameters estimates when difficulty was confounded with dimensionality. Ackerman reported that the confounding of difficulty and dimensionality effect was minimal and equal for both models. The relationships between the Logist parameter estimates and the true parameter values were similar to those reported by Way et al. (1988). Unfortunately Ackerman did not consider the relationship of the Logist item estimates and the means or sums of the true item parameter values as was reported by Way et al. However, he did report that the estimated ability parameters θ^* were best represented as the average of the true θ_1 and θ_2 values and that the correlation of a^* with a_1 approached that of a^* with a_2 as the correlation between the dimensions increased and that b^* was more highly correlated with b_1 than with b_2 for all relationships between the dimensions.

Summary

There is a large amount of support for the hypothesis that as the relationship between dimensions, in the case of IRT models, or the relationship between factors, in the case of factor analytical model, increases then so does the robustness of unidimensional IRT models to violations of the assumption of a single underlying trait. However, there is still one caveat, that is the appropriateness of the data generation methods. As has been already discussed with factor analytical procedures, there is the assumption of a linear relationship

between the trait being measured and the observed response. In the the case of MIRT models there is a lack of both estimation and confirmation procedures to ensure the validity of their use.

CHAPTER 3

Research Design and Methodology

The current study assessed the effects on the estimation of UIRT parameters when data sets violate the assumption of unidimensionality by exhibiting varying degrees of multidimensionality and correlations among the dimensions. The effects were assessed by generating data sets having two or three dimensions and correlations among the dimensions of 0.0, 0.3, 0.6, 0.95, and 0.99. Thus, ten data sets were generated, each representing a different combination of dimensionality and correlation among the dimensions.

The data were generated using a Fortran 77 computer program and IMSL subroutines. The suitability of the compensatory (CMIRT) and noncompensatory (NMIRT) multidimensional item response models used for generating the data were considered. Methods to ensure that the data sets did indeed contain the intended characteristics were employed.

The computer program Logist was employed to estimate the person and item parameters for the pseudo three parameter unidimensional IRT model (guessing parameter held constant at 0.2). The estimated parameters for each data set were compared to the parameters which were used in generating the data sets. The degree of congruity between the estimated parameters produced by Logist and the parameters inherent in the data sets was tested by examining the correlation between the IRT parameters, their means, and their sums. Mean square differences were also examined to determine the size of the congruence when dimensionality and correlation were varied. The results for the two dimensional data set are presented first and then the results for the three dimensional data set are presented. The two dimensional data set was generated by removing the third dimension from the three dimensional model.

Model Selection

Concerns regarding violations of the unidimensional assumption has been expressed for some time in the literature (Hattie, 1981, 1984, 1985; Harrison, 1986; Tucker, Humphreys, & Roznowski, 1986; McKinley, & Reckase 1982; Lord & Novick, 1968; Lord, 1980; Traub, 1983; Traub & Wolfe, 1981;). Hambleton et. al. (1978) assert that the testing of the assumption of unidimensionality is of a higher priority than the test of any of the remaining assumptions of unidimensional IRT models. They conclude that if the assumption of unidimensionality does not hold then the results of other tests are questionable (p. 487). There are two possible solutions. The first is to employ an index of unidimensionality to assess when unidimensional models are appropriate. The second is to develop multidimensional models which better reflect the underlying structure of the data regardless of the dimensionality. Indices of unidimensionality are at best questionable, with the possible exception of nonlinear factor analysis with respect to analysis of residuals (for a review see Hattie, 1985). Thus, the remaining solution is to develop models that relate the response of a subject to the number of dimensions involved in mediating the response.

A number of multidimensional models have been recorded in the IRT literature (Bock & Atkin, 1981; Doody-Bogan & Yen, 1983; Hattie, 1981; Mulaik, 1972; Rasch, 1961; Reckase, 1985; Samejima, 1974; Simpson, 1978; Whitely, 1980). Several categorizations for multidimensional models have been presented. One grouping categorized the models as being conjunctive, disjunctive and compensatory (Coombs, 1964; Combs & Kao 1954, cited in Hattie, 1984, pp. 55-56). More recently (Ackerman, 1989; Ansely & Forsyth, 1985; Way, Ansley, & Forsyth, 1988; Simpson, 1978), the terms compensatory and noncompensatory are presented as a classification system for the varying multidimensional models. Simpson (1978) referred to these models as being fully compensatory and partially compensatory respectively. However, the terms compensatory

and noncompensatory are widely used. Unfortunately the terms compensatory and noncompensatory may be somewhat misleading as to the nature of the difference between the two models. Compensatory models (Bock & Aitkin, 1981; Doody-Bogan & Yen, 1983; Hattie, 1981; Rasch, 1961) make the assumption that high ability on one dimension can compensate for low ability on another dimension. The noncompensatory models (Simpson, 1978; Whitely, 1980), assume that high ability on one dimension can only partially compensate for low ability on another dimension. For a complete discussion of the differences between these two classifications consult McKinley and Reckase (1982). A rather straightforward explanation is presented by Simpson (1978). The current study will focus on the compensatory (Doody-Bogan and Yen, 1983) and noncompensatory (Simpson, 1978) models developed as extensions of the unidimensional three parameter logistic model (Birnbaum, 1968).

Compensatory Item Response Models

The Compensatory Item Response (CIRM) model best demonstrates a factor analytical structure. The number of factors are a function of the way in which items differentially cluster. With respect to CIRM models, tests consisting of multiple dimensions have items clustering on each dimension. It is expected then that if the data fit a CIRM model then their factor structure which accurately reflect their dimensionality.

The three-parameter equation for $P_{ij}(\theta_{ih})$, the probability of person i correctly responding to item j , can be expressed as follows, where $D = 1.7$ (the scaling factor):

$$P_{ij}(\theta) = c_j + (1-c_j) \left[1 + e^{-D \sum_{h=1}^n a_{jh}(\theta_{ih} - b_{jh})} \right]^{-1}$$

Where:

P_{ij} = the probability of person i correctly answering item j

σ_{ih} = the *ability parameter* for person i for dimension h .

a_{jh} = the *multidimensional item discrimination parameter* for item j for dimension h .

b_{jh} = the *multidimensional item difficulty parameter* for item j for dimension h .

c_j = the *guessing parameter* for item j .

The defining character of the CIRM model is the denominator. The denominator is a function of summing across the varying dimensions. Therefore, as mentioned previously, if an individual is deficient on one dimension then high ability on another dimension can compensate for the deficiency.

Noncompensatory Item Response Models

As with the CIRT model, the defining characteristic of the Noncompensatory Item Response (NIRT) model is the denominator of the function $f(P_{ij})$. Here it can be seen that the relationship between the different dimensions is a multiplicative one. As previously mentioned an NIRT model allows for partial compensation by having high ability on one dimension making up for relatively low ability on another dimension. However, if on any one dimension, the ability needed to answer the question is zero, then no compensation is possible. For example, if an individual had an infinite amount of ability on $h-1$ dimensions and zero ability of the g th dimension then $P(\theta) = 0.0$ (where guessing is not a factor). This contrasts sharply with the CIRT models where the relationship between the dimensions is a summative one, thus, allowing zero ability on one dimension to be compensated by high ability on other dimensions. The theoretical weakness of fully compensatory models led Symphon (1978) to argue against the practicality of their use with measures of ability.

The three-parameter equation for $P_{ij}(\theta_i)$ can be expressed as follows:

$$P_{ij}(\theta) = c_j + (1-c_j) \prod_{h=1}^n [1 + e^{-Da_{jh}(\theta_{ih}-b_{jh})}]^{-1}$$

Where:

$D = 1.7$ the scaling factor

P_{ij} = the probability of person i correctly answering item j

π_h indicates multiplication across the dimensions $h=1,2,\dots,n$

θ_{ih} = the ability parameter for person i for dimension h .

a_{jh} = the multidimensional item discrimination parameter for item j for dimension h .

b_{jh} = the multidimensional item difficulty parameter for item j for dimension h .

c_j = the guessing parameter for item j .

In summary, multidimensional models in general fall into one of two classifications, compensatory or noncompensatory. With respect to the compensatory models it is reasonable to expect that if one had a large amount of ability on one dimension then this would make up for a small amount of ability on another dimension. Suppose a paper and pencil test consisting of simple addition and subtraction were given to an individual who had demonstrated the ability to add and subtract, through concrete examples. Suppose further that this individual could not read. Thus, one would expect that this individual would score zero on the test or at least have a very low mark in the case where guessing is a factor. The compensatory model does not adequately allow for such a condition. However, the Simpson (1978) model does allow for these types of boundary conditions, as well as allowing for compensation to occur when a large amount of ability exist on one dimension and a small amount of ability exist on another dimension. It can be seen that the categorization of these two models is somewhat unfortunate. More correctly they are both compensatory models. The Doody-Bogan and Yen (1983) model, with its summative

nature, is a more limited example and the Simpson (1978) model, with its multiplicative nature, is a more encompassing example. On the basis of the above discussion Simpson's (1978) model was selected to generate the multidimensional pseudo-responses.

Parameter Selection

Parameter selection operates at three levels. At the first level are the theoretical considerations, at the second level are the empirical considerations and at the third level are the practical considerations. Theoretically, the item and trait parameters are bounded by plus and minus infinity for the two parameter logistic model (Lord & Novick, 1968). For the three parameter model the added item parameter c is bounded by zero and one.

Empirically, the a parameter is positive and ranges generally from 0.5 to 2.0 with a mean value of approximately 1.0 and a standard deviation of approximately 0.4 (Lord, 1968, 1980; Ree, 1979; Ross, 1966). The b parameter generally ranges from plus 2 to minus 2.0 with a mean approximately 0.5 and a standard deviation of approximately 1.0 (Hattie, 1984; Lord, 1968). The ϕ parameter ranges generally from plus 3.0 to minus 3.0 with a mean approximately 0.0 and a standard deviation of approximately 1.0 (Hulin, Drasgow, and Parson, 1982; Lord, 1968). The c parameter has been reported to range from 0.04 to 0.20 with a mean 0.16 and standard deviation of 0.01 (Lord, 1968) and from 0.09 to 0.35 with a mean of 0.2 and standard deviation of 0.05 (Ree, 1979). In the current study a pseudo three parameter model was used to both generate the data and to estimate the parameters (c parameter held constant at 0.2). The c parameter often does not converge during estimation and this tends to destabilize the entire estimation procedure. Since the usefulness of the c parameter for the results of this study are negligible and the possibility of negative impacts high, the c parameter was held constant. Parameter selection becomes a process of trial and error. For the current study item and trait parameters were generated based on the procedures outlined in Ansley and Forsyth (1985).

Data Generation

To produce the response matrix, a computer program (Non-Comp) written in Fortran 77 was developed. The program utilizes seven International Mathematical and Statistical Libraries (IMSL) subroutines to generate random numbers and to control the relationships between the dimensions in the model (International Mathematical and Statistical Libraries, 1987). The accuracy of the generated parameters was checked using statistical procedures provided through the statistical program SPSS (Statistical Package for the Social Sciences, 1988). All data generation and statistical analysis were conducted using the AMDAHL main frame computer at the University of Alberta.

Currently Non-Comp is capable of simulating response data for up to 100 items and 2000 subjects on 5 dimensions. The response matrix consist of 0's and 1's, where a "0" indicates an incorrect response and a "1" indicates a correct response. Input to the program varies depending on the particular design required. In general the first record contains the title for the current simulation. The second record contains the type of model (compensatory or noncompensatory), the distribution for the difficulty parameter (uniform or normal), the number of dimensions (1 to 5), the number of items (1 to 100), the number of subjects (1 to 2000) and the seed number for the random number generation. The third through fifth records contain the variance-covariance matrices used to control the correlation between the dimensions in the model. These variance-covariance matrices are related to the ability, difficulty and discrimination parameters, respectively.

The ability parameters (A) are pseudo-random numbers generated from a multivariate normal distribution.

$${}_N A_n \sim N(0,1)$$

Where:

N is the number of subjects (2000 subjects in the current study).

n is the number of dimensions (3 dimensions in the curent study).

The N vectors of ability parameters, representing the n dimensions, are then orthogonalized (${}_N F_n$) and postmultiplied by an upper triangular factorization (Cholesky (${}_n C_n$)) of the variance/covariance matrix (${}_n V_n$).

$${}_N F_n = N(0,1); \text{ Covariance} = 0$$

$${}_n V_n = {}_n C_n {}_n C_n'$$

$${}_N A_n' = {}_N F_n {}_n C_n$$

The distribution of parameters within each dimension were then rescaled to the desired mean and standard deviation. For the current study the mean and standard deviation are zero and one respectively.

The difficulty parameters may be generated from either a uniform or a multivariate normal distribution. For the current study the difficulty parameters were generated from a multivariate normal distribution. The procedures for controlling the correlations between the dimensions were the same as those employed with the ability parameters. Unlike the ability parameters the distributional characteristics of the difficulty parameters were differentially set across the three dimensions. For the first dimension the mean was set at 0.33 and the standard deviation was set at 0.82, while for the second dimension the mean was -1.03 and the standard deviation was 0.82 (Ansley & Forsyth, 1985). The rationale for setting the second dimension such that its mean difficulty was lower than that for the first dimension was an intuitive one. If one imagines a mathematics test consisting of word problems then one might expect that the level of reading difficulty would be somewhat lower than the level of mathematical difficulty. In the current study a third dimension was added and the same rationale was extended to this dimension. Consequently the mean was set to -1.55 and the standard deviation was 0.82.

The discrimination parameters were generated from a uniform distribution using the same methods mentioned above. The means for the three dimensions were set at 1.23, 0.49 and 0.25 respectively. The standard deviations were set at 0.34, 0.11 and 0.11. A similar rationale as that mentioned for the difficulty parameters can be applied for selecting

the values for the discrimination parameters. If the mathematics test had been designed to distinguish between students primarily on their level of mathematics ability and not on their reading ability, then it is reasonable to assume that items would be designed such that they discriminate more highly on their primary dimensions.

Simulation Model

Once the item and trait parameters had been generated, the probability of person i responding to item j correctly was calculated for each person and item in the simulation, using the model developed by Sympson (1968).

$$P_{ij}(\theta) = c_j + (1 - c_j) \prod_{h=1}^n [1 + e^{-D a_{jh}(\theta_{ih} - b_{jh})}]^{-1}$$

In the current study this required generating 2000 by 60 response probabilities (p_{ij}) for each of the 10 conditions. Subsequently, each probability was compared to a threshold value (r_{ij}) generated from a uniform distribution $U(0,1)$. If the probability of the response was greater than or equal to the generated threshold value the response (x_{ij}) was set to "1"; otherwise the response was set to "0", that is

$$\begin{aligned} x_{ij} &= 1 \text{ if } p_{ij} \geq r_{ij} \\ x_{ij} &= 0 \text{ if } p_{ij} < r_{ij}. \end{aligned}$$


CHAPTER 4

Findings and Analysis

Two Dimensional Data Structure

Data Confirmation

A consideration of prime importance was to ensure that the characteristics intended to be incorporated into the generated data sets did indeed exist. To assess whether the data set generated exhibited the intended characteristics a number of statistics relevant to item analysis were produced (see Table 1). The values shown are consistent with test statistics reported in the literature (Ansley & Forsyth, 1985).

In Table 2 are the results of a principal axes factor analysis with varimax and promax rotations. The initial estimates of the communalities were made equal to the squared multiple correlations. The data were generated on the basis of a two dimensional model with varying degrees of correlation between the dimensions. In all, five separate factor analyses were performed. The second column gives the correlation between the factors for a two factor solution. The general trend was that the correlation between the factors increased as the correlation between the dimensions in the model increased. The third and fourth columns give the percentage of common variance accounted for by the first and second factors in both the varimax and promax rotations. As the correlation between the factors increased the general trend was for more of the variance to be accounted for by the first factor and less of the variance to be accounted for by the second factor. This trend held for both forms of rotation. Despite this fact, the results tend to support the conclusion that a two dimensional structure did underlie the data as was intended. These results were consistent with the findings of Ansley and Forsythe (1985).

Table 3 provides further evidence for a two dimensional data structure. In columns 1 through 3 are the ratios of the first and second, second and third, and third and fourth eigenvalues for the two dimensional data sets. The relative size of the first ratio to the

Table 1
Descriptive Statistics for Two dimensional Data

Corr ϕ_1, ϕ_2	λ_1/λ_2	α	Mean p	Range p	Range Bis	Mean Score	Range Score	Skew	Kurt
0.00	07.96	0.87	0.50	.24-.72	.24-.55	30.14	6-58	0.23	-0.62
0.30	10.13	0.89	0.51	.27-.73	.27-.58	30.86	7-57	0.17	-.072
0.60	11.76	0.91	0.53	.26-.73	.31-.60	31.48	7-60	0.16	-.085
0.90	12.61	0.92	0.53	.30-.74	.34-.63	32.02	6-60	0.20	-0.89
0.95	12.70	0.92	0.54	.27-.74	.36-.63	32.12	6-60	0.22	-.095

Descriptive statistics, ratio of first and second eigenvalues (λ_1/λ_2), test reliability (α), test difficulty (p), item-total biserial correlation (Bis), for two dimensional data with number of subjects equal to 2000 and number of items equal to 60.

Table 2

Factor Analysis of Two Dimensional Data

Corr ϕ_1, ϕ_2	Corr f_1, f_2	Varimax % f_1	% f_2	%T	Promax % f_1	% f_2	%T
0.00	0.730	51	37	88	54	35	89
0.30	0.773	58	38	96	62	35	97
0.60	0.776	57	35	92	62	30	92
0.90	0.780	58	36	94	63	31	94
0.95	0.780	58	36	94	63	31	94

Principal axes with varimax and promax rotation. Square multiple correlations substituted in the main diagonal. The correlation between ability dimension 1 and 2 $\rho(\phi_1, \phi_2)$. The correlation between the factors $\rho(f_1, f_2)$. The percentage of the total variance accounted for by factor 1 % f_1 and factor 2 % f_2 and the total variance explained by f_1 and f_2 for both a varimax and promax rotation are presented.

second and third would indicate a strong first factor. The same is true of the second ratio, which would seem to indicate a moderate second factor. However, the third ratio is not much different from a value of one, indicating roots of approximately the same size. It should be noted that this trend held for the remaining eigenvalues:

A final method employed to verify the integrity of the data was that of multidimensional scaling, in which an alternating least squares method (ALSCAL) is employed (Takane, Forest, Young, & Leeuw, 1977). Table 4 provides the stress values based on Kruskal's stress formula number 1 (Kruskal & Wish, 1978). Also provided are the R-Squared values. Because of the large computational expense to run such a procedure only the results for two dimensional data when the dimensions are orthogonal are presented. The first row indicates the amount of stress to fit a five dimensional structure to the data. Subsequent rows are the 4, 3, 2 and 1 dimensional solutions. As can be seen, the maximum amount of stress is attained when the shift is made from a two dimensional to a one dimensional solution. These results are in keeping with what would be expected for data of this nature and is further confirming evidence that the data generation procedures are successfully reproducing the type of structure intended. There is one word of caution, however, some researchers, Hattie (1985) in particular, feel that linear factor analytical techniques do not provide appropriate indices of dimensionality. Further, little is known about the predictive power of multidimensional scaling as an index of dimensionality. However, linear factor analysis has been widely used as guide to dimensionality and does provide some useful insight into the nature of the data's underlying structure.

Data Analysis

Estimation of the item and trait parameters was derived using the computer program Logist (Wingersky, Barton, & Lord, 1982). For the current study a pseudo three parameter model was selected (guessing parameter set to 0.2) and the number of choices per item was set to five. All other options were left at the default settings.

Table 3

Eigenvalue Ratios for the Two Dimensional Data

$p(\theta_1, \theta_2)$	$p(f_1, f_2)$	λ_1/λ_2	λ_2/λ_3	λ_3/λ_4
0.00	0.730	7.96	2.17	1.13
0.30	0.773	10.13	2.11	1.32
0.60	0.776	11.76	2.28	1.19
0.90	0.780	12.61	2.79	1.06
0.95	0.780	12.70	2.80	1.08

Ratio of first and second eigenvalues λ_1/λ_2 , second and third eigenvalues λ_2/λ_3 , and the third and fourth eigenvalues λ_3/λ_4 at each correlation between the dimensions $p(\theta_1, \theta_2)$.

Table 4

Multidimensional Scaling for Two Dimensional Data

Dim.	Stress	RSQ	dS	dRSQ
5	0.09	0.95		
4	0.10	0.95	0.01	0.00
3	0.12	0.94	0.02	0.01
2	0.14	0.92	0.02	0.02
1	0.20	0.88	0.06	0.04

Stress and squared correlation (RSQ) in distances. RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula number 1. dS is the change in Stress and dRSQ is the change in RSQ.

Once the estimated model parameters were ascertained, the analysis consisted of correlating these estimates with the true parameter values and their means, which were computed by summing across the dimensions and dividing by the number of dimensions. Further, average absolute differences (AAD) were computed to aid in ascertaining the discrepancy between the estimated and true parameters. Thus, for the ability parameters the AAD's were of the form:

$$AAD_{\theta_h} = \frac{\sum_{i=1}^N |\theta_{ih} - \theta_i^*|}{N}$$

where :

θ_{ih} ($h = 1, 2$) is the true ability parameter for person i for dimensions 1 and

2.

θ_i^* is the estimated ability parameter for person i .

N is the number of examinees

where the summation is over the N individuals for a given dimension.

The form of the AAD for the item parameters is similar in nature

$$AAD_{x_h} = \frac{\sum_{j=1}^K |x_{jh} - x_j^*|}{K}$$

Where

x_{jh} ($h = 1, 2$) is the true difficulty or discrimination parameter for

dimensions 1 and 2.

x_j^* is the estimated difficulty or discrimination parameter for item j

K is the number of items

where the summation is over the K items for a given dimension.

Results

Ability. The correlations of θ^* with θ_1 , θ_2 , and the mean of θ_1 and θ_2 (M_θ) at each of the five values of ρ (θ_1, θ_2) are presented in Table 5. As ρ (θ_1, θ_2) increased the correlation of θ^* with θ_1 , θ_2 , and M_θ increased. However, θ^* had a strong association with M_θ at all levels of ρ (θ_1, θ_2). Of special interest was the relative change in the correlation of θ^* with the remaining parameters as ρ (θ_1, θ_2) moves from 0.0 to 0.95. For ρ (θ^*, θ_1) the range was 0.09 while for ρ (θ^*, M_θ) the range was 0.05, however, for ρ (θ^*, θ_2) the range was 0.44. It should be noted that, although the relationship of θ^* with the average of θ_1 and θ_2 was the overall the strongest, it was not substantially different from that of ρ (θ^*, θ_1).

In Table 5 are also given the average absolute differences between the estimated ability parameters and the true ability parameters and their mean summed across the dimensions. Again, there was a clearly a smaller discrepancy between θ^* and θ_1 than there was between θ^* and θ_2 however this difference, as before, becomes smaller as ρ (θ_1, θ_2) increased. The smallest average absolute difference overall was associated with M_θ .

These results parallel those presented by Ansley and Forsyth (1985). Both studies clearly indicate that the estimated ability parameter for low values of ρ (θ_1, θ_2) are most strongly associated with M_θ , however, this association is not significantly different than their association with θ_1 . At high levels of ρ (θ_1, θ_2), θ^* is equally associated with θ_1 , θ_2 , and M_θ .

Discrimination. The correlations of a^* with a_1 , a_2 , and the mean of a_1 and a_2 (M_a) and the average absolute differences at each of the five levels of ρ (θ_1, θ_2) are presented in Table 6. Columns 2 through 4 clearly indicate that the magnitude of ρ (θ_1, θ_2) did not substantially affect their relationships. However, a^* associated more strongly with a_1 , than it did with a_2 for all levels of ρ (θ_1, θ_2). Further, a^* associated just as strongly with a_1 as it did with M_a .

Table 5
Correlations and Absolute Mean Differences Between the Estimated and True Ability
Parameters for Two Dimensional Data

$\rho(\theta_1, \theta_2)$	$\rho(\theta^*, \theta_1)$	$\rho(\theta^*, \theta_2)$	$\rho(\theta^*, M_\theta)$	AAD _{θ_1}	AAD _{θ_2}	AAD _{M_θ}
0.00	0.74	0.38	0.79	0.56	0.93	0.48
0.30	0.78	0.55	0.82	0.51	0.80	0.44
0.60	0.81	0.64	0.84	0.44	0.63	0.38
0.90	0.83	0.81	0.84	0.39	0.45	0.36
0.95	0.83	0.82	0.84	0.37	0.41	0.36

Correlations of the estimated ability parameter (θ^*) with the true ability parameters (θ_1, θ_2) and their mean (M_θ) for dimensions 1 and 2. The average absolute differences are AAD _{θ_1} , AAD _{θ_2} , and AAD _{M_θ}

Table 6
Correlations and Absolute Mean Differences Between the Estimated and True
Discrimination Parameters for Two dimensional Data

$\rho(\theta_1, \theta_2)$	$\rho(a^*, a_1)$	$\rho(a^*, a_2)$	$\rho(a^*, M_a)$	AAD _{a_1}	AAD _{a_2}	AAD _{M_a}
0.00	0.93	-0.22	0.90	0.45	0.32	0.12
0.30	0.94	-0.22	0.91	0.37	0.40	0.10
0.60	0.94	-0.23	0.91	0.28	0.47	0.13
0.90	0.95	-0.21	0.92	0.20	0.55	0.19
0.95	0.95	-0.20	0.93	0.18	0.57	0.20

Correlations of the estimated discrimination parameter (a^*) with the true discrimination parameters (a_1, a_2) and their mean (M_a) for dimension 1 and 2. The average absolute differences are AAD _{a_1} , AAD _{a_2} , and AAD _{M_a}

With respect to the AAD's it can be seen that there was an interaction between the levels of the true discrimination parameters and the levels of ρ (θ_1, θ_2). At low levels of ρ (θ_1, θ_2) the $AADa_2$ was smaller than $AADa_1$, while at large values of ρ (θ_1, θ_2) the $AADa_2$ was larger than $AADa_1$. This same relationship was seen between a_1 and M_a however, it was not as pronounced.

In summary, a^* appeared to be most highly related to M_a . This is in keeping with results reported elsewhere in the literature (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988; but see Ackerman, 1989). However, these results may in part be an artifact of the way in which the parameters were generated. Recall that the standard deviation of a_1 was set at 0.34 while the standard deviation for a_2 was set at 0.11. Therefore, one would expect this larger variance to contribute in part to the larger correlation. One striking difference between the results in the current study and those found in the Ansley and Forsyth (1985) study is the magnitude of the correlations between the estimated parameters and the true parameters. Ansley and Forsyth (1985) reported correlations ranging from 0.47 to 0.64 for $\rho(a^*, a_1)$ and 0.02 to -0.05 for $\rho(a^*, a_2)$ while $\rho(a^*, M_a)$ ranged from 0.50 to 0.65. In the current study $\rho(a^*, a_1, M_a)$ were greater than 0.90 for all levels of $\rho(\theta_1, \theta_2)$, and $\rho(a^*, a_2)$ ranged from -0.22 to -0.20 as $\rho(\theta_1, \theta_2)$ increased. It is not clear at this time why that was the case.

Difficulty. The correlations of b^* with b_1, b_2 , and the mean of b_1 and b_2 (M_b), and the average absolute differences at each of the five levels of ρ (θ_1, θ_2) are presented in Table 7. With respect to the correlations, it can be clearly seen that as ρ (θ_1, θ_2) increased $\rho(b^*, b_1)$ increased and $\rho(b^*, b_2)$ decreased, indicating a much stronger relationship between estimated difficulty parameters and the first dimension than between estimated difficulties and the second dimension. Further, as was the case with the ability parameters and the discrimination parameters the strongest relationship was with the average of the two dimensions, however, $\rho(b^*, M_b)$ is not substantially greater than $\rho(b^*, b_1)$ for all levels of ρ (θ_1, θ_2).

Table 7 also indicates that all AAD's decreased as $\rho(\theta_1, \theta_2)$ increased. Further, the smallest AAD was between b^* and b_1 , indicating that b^* was a better estimate of b_1 . An interesting observation was that b^* consistently over estimated b_1 , but b^* approached b_1 as $\rho(\theta_1, \theta_2)$ increased.

Three Dimensional Data Structure

Data Confirmation

As with the two dimensional data structure, it was once again necessary to ensure that the intended characteristics had been incorporated into the generated data sets. The same procedures used to generate the two dimensional data were repeated in generating the three dimensional data. To assess the integrity of the data, in the three dimensional case, a number of statistics relevant to classical item analysis were produced (see Table 8). Overall it was apparent that the test difficulty had increased ($0.42 \leq \text{Mean}(p) \leq 0.45$) compared with those results reported in Table 1 ($0.50 \leq \text{Mean}(p) \leq 0.54$). Further evidence of an increased test difficulty was found in the reduction of the mean raw score for the three dimensional data structure ($25.17 \leq \text{Mean}(x) \leq 27.22$), relative to the mean of the raw score for the two dimensional data structure ($30.14 \leq \text{Mean}(x) \leq 32.12$). Of special interest was the overall increase in the ratio of λ_1/λ_2 with an increase in the dimensionality from 2 dimensions ($7.96 \leq \lambda_1/\lambda_2 \leq 12.70$) to three dimensions ($9.42 \leq \lambda_1/\lambda_2 \leq 16.39$).

In Table 9 are the results of a principal axes factor analysis with a varimax and a promax rotation. The initial estimates of the communalities were made equal to the squared multiple correlations. The data were generated on the basis of a three dimensional model with varying degrees of correlation between the dimensions. Five separate factor analyses were employed. The second through fourth columns give the correlations between the factors for a three factor solution. The general trend was that the correlation between the factors increased as the the correlation between the dimension in the model increased. The fifth through twelfth columns give the percentage of common variance accounted for in the

Table 7
Correlations and Absolute Mean Differences Between the Estimated and True Difficulty Parameters for Two Dimensional Data

$\rho(\theta_1, \theta_2)$	$\rho(b^*, b_1)$	$\rho(b^*, b_2)$	$\rho(b^*, M_b)$	AAD _{b1}	AAD _{b2}	AAD _{mb}
0.00	0.900	0.700	0.96	0.80	1.50	1.15
0.30	0.910	0.690	0.96	0.69	1.39	1.04
0.60	0.917	0.680	0.96	0.63	1.33	0.98
0.90	0.921	0.668	0.96	0.56	1.26	0.91
0.95	0.922	0.664	0.96	0.549	1.25	0.89

Correlations of the estimated difficulty parameter (b^*) with the true difficulty parameters (b_1, b_2) and their mean (M_b) for dimensions 1 and 2. The average absolute differences are AAD_{b1}, AAD_{b2}, and AAD_{mb}.

Table 8

Descriptive Statistics for Three Dimensional Data

Corr $\theta_1\theta_2\theta_3$	λ_1/λ_2	α	Mean p	Range p	Range Bis	Mean Score	Range Score	Skew	Kurt
0.00	9.42	0.77	0.42	.24-.64	.18-.52	25.17	6-49	0.123	-0.446
0.30	11.31	0.81	0.43	.24-.64	.21-.49	25.73	4-49	0.197	-0.550
0.60	13.35	0.84	0.44	.25-.65	.22-.53	26.43	4-52	0.240	-0.615
0.90	15.93	0.87	0.45	.25-.66	.24-.54	27.07	4-55	0.249	-0.698
0.95	16.39	0.87	0.45	.25-.67	.24-.55	27.22	4-55	0.222	-0.745

Descriptive statistics, ratio of first and second eigenvalues (λ_1/λ_2), test reliability (α), test difficulty (p), item-total biserial correlation (Bis), for three dimensional data with number of subjects equal to 2000 and number of items equal to 60.

three factor solution with respect to both the varimax and promax rotations. Unlike the results found for the two dimensional case (see Table 2), where the largest portion of the variance was contributed by the first factor, it was apparent that each of the three factors extracted from the three dimension data set, were substantially contributing to the overall variance observed in the data. However, the trend was that the first factor contributed the largest portion while the second factor contributed the next largest portion for all values of $\rho(\phi_1, \phi_2, \phi_3)$. Further, it is evident that, as $\rho(\phi_1, \phi_2, \phi_3)$ increased the percent of the total variance accounted for by the three factor solution increased accordingly. This suggests that the data space was becoming more unidimensional in its nature as the relationship between the dimensions increased.

Table 10 provides further evidence about dimensionality of the data structure. In columns 4 through 6 are the ratios of the first and second, second and third, and third and fourth eigenvalues for the three dimensional data sets. The relative size of the first ratio to the second and third would indicate a strong first factor. However, subsequent ratios appear to indicate that no further factors can be inferred from the output. It appears that the effect of adding a third dimension to the data structure is that of causing the data to take on a structure which is factor analytically unidimensional.

A final method employed to verify the integrity of the data is that of multidimensional scaling. Table 11 provides the stress values based on Kruskal's stress formula number 1. Also provided are the R-Squared values. Once again, because of the large expense to run such a procedure, only the results for three dimensional data when the dimensions are orthogonal are presented. The first row indicates the amount of stress to fit a five dimensional structure to the data. Subsequent rows are the 4, 3, 2 and 1 dimensional solutions. Unlike the two dimensional data set (see Table 4), no clear indication as to the dimensionality of the data set was provided by the multidimensional scaling technique. It would appear, at least in this instance, that multidimensional scaling is of limited value as an index of dimensionality.

Table 9

Factor Analysis of Three Dimensional Data

Corr	Corr			Varimax				Promax			
$\vartheta_1, \vartheta_2, \vartheta_3$	f_1, f_2	f_1, f_3	f_2, f_3	% f_1	% f_2	% f_3	%T	% f_1	% f_2	% f_3	%T
0.00	0.693	0.664	0.623	35	30	29	94	37	29	29	95
0.30	0.720	0.616	0.603	36	35	26	97	39	36	22	97
0.60	0.731	0.702	0.653	40	32	27	99	43	31	24	98
0.90	0.774	0.776	0.768	37	32	29	98	38	31	29	98
0.95	0.780	0.750	0.770	36	35	29	100	37	35	27	99

Principal axes with varimax and promax rotation, and square multiple correlations substituted in the main diagonal. The correlation between dimension 1, 2, and 3 $\rho(\vartheta_1, \vartheta_2, \vartheta_3)$. The correlation between the factors $\rho(f_1, f_2, f_3)$. The percentage of the total variance accounted for by factor 1 (% f_1) and factor 2 (% f_2) and factor 3 (% f_3) and the total variance explained (%T) by f_1, f_2 and f_3 for both a varimax and promax rotation.

Table 10

Eigenvalue Ratios for the Three Dimensional Data

$\rho(\vartheta_1, \vartheta_2, \vartheta_3)$	$\rho(f_1, f_2)$	$\rho(f_1, f_3)$	$\rho(f_2, f_3)$	λ_1/λ_2	λ_2/λ_3	λ_3/λ_4
0.00	0.693	0.664	0.623	9.42	1.09	1.05
0.30	0.720	0.616	0.603	11.31	1.08	1.13
0.60	0.731	0.702	0.653	13.35	1.23	1.08
0.90	0.774	0.776	0.768	15.93	1.16	1.10
0.95	0.780	0.750	0.770	16.39	1.18	1.10

Ratio of first and second eigenvalues λ_1/λ_2 , second and third eigenvalues λ_2/λ_3 , and the third and fourth eigenvalues λ_3/λ_4 at each correlation between the dimensions $\rho(\vartheta_1, \vartheta_2, \vartheta_3)$.

Table 11

Multidimensional Scaling for Three Dimensional Data

Dim.	Stress	RSQ	dS	dRSQ
5	0.13	0.88		
4	0.15	0.86	0.02	0.02
3	0.18	0.83	0.03	0.03
2	0.22	0.80	0.04	0.03
1	0.30	0.75	0.08	0.05

Stress and squared correlation (RSQ) in distances. RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula number 1. dS change in Stress dRSQ change in RSQ.

Data Analysis

As with the analysis of the two dimensional data structure, estimation of the item and trait parameters were derived using the computer program Logist (Wingersky, Barton, & Lord, 1982). Once again a pseudo three parameter model was selected (guessing parameter set to 0.2) and the number of choices per item was set to five. All other options were left at the default settings.

Once the estimated model parameters were ascertained, analysis consisted of correlating these estimates with the true parameter values and their means, which were computed by summing across the dimensions and dividing by the number of dimensions. Further, average absolute differences (AAD) were computed to aid in ascertaining the relationship between the estimated and true parameters. Thus, for the ability parameters the AAD's were of the form:

$$AAD_{\theta_h} = \sum_{i=1}^N \frac{|\theta_{ih} - \theta_i^*|}{N}$$

where :

θ_{ih} ($h = 1, 2$ and 3) is the true ability parameter for person i for dimensions 1 to 3.

θ_i^* is the estimated ability parameter for person i .

N is the number of examinees

where the summation is over the N people for a given dimension.

The form of the AAD for the item parameters is similar in nature:

$$AAD_{x_h} = \sum_{j=1}^K \frac{|\bar{x}_{jh} - \bar{x}_j^*|}{K}$$

Where

x_{jh} ($h = 1, 2$ and 3) is the true difficulty or discrimination parameter for dimensions 1 to 3.

x_j^* is the estimated difficulty or discrimination parameter for item j

K is the number of items

where the summation is over the K items for a given dimension.

Results

Ability. The correlations of ϕ^* with ϕ_1, ϕ_2, ϕ_3 , and the mean of ϕ_1, ϕ_2 , and ϕ_3 (M_ϕ) at each of the five levels of ρ (ϕ_1, ϕ_2, ϕ_3) are presented in Table 12. As $\rho(\phi_1, \phi_2, \phi_3)$ increased the correlation of ϕ^* with ϕ_1, ϕ_2 , and M_ϕ increased. Unlike the two dimensional case, where ϕ^* had its strongest association with M_ϕ at all levels of ρ (ϕ_1, ϕ_2), there appeared to be no real difference in the relationship between ϕ^* and ϕ_1 , or ϕ^* and M_ϕ . Of special interest was the relative change in the correlation of ϕ^* with the remaining parameters as ρ (ϕ_1, ϕ_2, ϕ_3) moved from 0.0 to 0.95. For $\rho(\phi^*, \phi_1)$ the range was 0.09 while for $\rho(\phi^*, M_\phi)$ the range was 0.13; however, for $\rho(\phi^*, \phi_2)$ this range was 0.46 and for $\rho(\phi^*, \phi_3)$ the range was 0.61.

In Table 12 are also given the average absolute differences between the estimated ability parameters and the true ability parameters and their mean summed across the dimensions. Again there was clearly a stronger relationship between ϕ^* and ϕ_1 than was found between ϕ^* and ϕ_2 or ϕ^* and ϕ_3 , however this difference, as before, became small as $\rho(\phi_1, \phi_2, \phi_3)$ increased. The smallest average absolute difference overall was associated with M_ϕ .

These results parallel those presented by Ansley and Forsyth (1985) and those presented for the two dimensional data in the current study. Both studies clearly indicate that the estimated ability parameter for low values of ρ (ϕ_1, ϕ_2, ϕ_3) are most strongly associated with M_ϕ , however, this association is not substantially different from their

Table 12
Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data

$\rho(\theta_1, \theta_2, \theta_3)$	$\rho(\theta^*, \theta_1)$	$\rho(\theta^*, \theta_2)$	$\rho(\theta^*, \theta_3)$	$\rho(\theta^*, M_\theta)$	AAD _{θ_1}	AAD _{θ_2}	AAD _{θ_3}	AAD _{m_θ}
0.00	0.68	0.29	0.14	0.64	0.75	1.09	1.22	0.75
0.30	0.68	0.46	0.37	0.69	0.71	0.96	1.04	0.67
0.60	0.73	0.61	0.55	0.74	0.67	0.82	0.89	0.63
0.90	0.76	0.73	0.72	0.76	0.61	0.67	0.69	0.60
0.95	0.77	0.75	0.75	0.77	0.59	0.62	0.63	0.58

Correlations of the estimated ability parameter (θ^*) with the true ability parameters ($\theta_1, \theta_2, \theta_3$) and their mean (M_θ) for dimensions 1, 2 and 3. The average absolute differences are AAD _{θ_1} , AAD _{θ_2} , AAD _{θ_3} , and AAD _{m_θ}

Table 13
Correlations and Absolute Mean Differences Between the Estimated and True Discrimination Parameters for Three Dimensional Data

$\rho(a_1, a_2, a_3)$	$\rho(a^*, a_1)$	$\rho(a^*, a_2)$	$\rho(a^*, a_3)$	$\rho(a^*, M_a)$	AAD _{a_1}	AAD _{a_2}	AAD _{a_3}	AAD _{m_a}
0.00	0.09	-0.06	0.32	0.19	0.62	0.22	0.37	0.18
0.30	0.11	-0.07	0.25	0.19	0.55	0.28	0.45	0.22
0.60	0.08	-0.06	0.32	0.18	0.52	0.31	0.50	0.23
0.90	0.03	-0.06	0.34	0.14	0.49	0.36	0.56	0.27
0.95	0.03	-0.06	0.32	0.13	0.49	0.37	0.57	0.28

Correlations of the estimated discrimination parameter (a^*) with the true discrimination parameters (a_1, a_2, a_3) and their mean (M_a) for dimension 1, 2, and 3. The average absolute differences are AAD _{a_1} , AAD _{a_2} , AAD _{a_3} , and AAD _{m_a}

association with θ_1 . At high levels of ρ ($\theta_1, \theta_2, \theta_3$), θ^* is equally associated with θ_1 , θ_2 , θ_3 and M_θ .

Discrimination. The correlations of a^* with a_1, a_2, a_3 and the mean of a_1, a_2 , and a_3 (M_a) and the average absolute differences at each of the five levels of ρ ($\theta_1, \theta_2, \theta_3$) are presented in Table 13. Columns 2 through 5 clearly indicate that the magnitude of ρ ($\theta_1, \theta_2, \theta_3$) did not significantly affect the relationship of a^* with a_1, a_2, a_3 and M_a . However, a^* associated more strongly with a_3 , than it did with a_1, a_2 or M_a for all levels of ρ ($\theta_1, \theta_2, \theta_3$). This high association between the estimated discrimination parameter and the true discrimination parameter of the third dimension contrasts sharply with the results reported for the two dimensional data structure, where a^* was most clearly related to the first dimension (see Table 6).

With respect to the AAD's, it can be seen that as ρ ($\theta_1, \theta_2, \theta_3$) increased AAD_{a_1} decreased, while all other AAD's increased in magnitude. However, overall AAD_{M_a} was the smallest for all AAD values of ρ ($\theta_1, \theta_2, \theta_3$).

In summary, a^* appeared to be most highly related to a_3 , however, the smallest AAD was associated with M_a . These results prove to be inconclusive and do not lend themselves to any straightforward interpretation of a^* 's relationship to the true parameters. One possible explanation may lie in the nature of the compensatory model. Recall that its distinguishing feature is the multiplicative relationship which appears in the denominator of the equation that describes the model. Recall also that in the initial generation of the discrimination parameters, the mean and standard deviation of the discrimination parameter for the third dimension was set to be somewhat lower than for the remaining dimensions. Therefore, because of the multiplicative nature of the model, the smallest value in fact becomes an upper bound on the overall probability. Thus, it may be that the higher correlation between a^* and a_3 was a result of the nature of the model.

Difficulty. The correlations of b^* with b_1, b_2, b_3 , and with the mean of b_1, b_2 , and b_3 , (M_b) and the average absolute differences at each of the five levels of

Table 14
*Correlations and Absolute Mean Differences Between the Estimated and True Difficulty
 Parameters for Three Dimensional Data*

$\rho(\theta_1, \theta_2, \theta_3)$	$\rho(b^*, b_1)$	$\rho(b^*, b_2)$	$\rho(b^*, b_3)$	$\rho(b^*, M_b)$	AAD _{b1}	AAD _{b2}	AAD _{b3}	AAD _{Mb}
0.00	0.76	0.67	0.62	0.89	1.47	2.17	3.14	2.26
0.30	0.71	0.67	0.66	0.89	1.30	2.00	2.97	2.09
0.60	0.76	0.66	0.63	0.89	1.16	1.86	2.83	1.95
0.90	0.77	0.66	0.62	0.89	1.05	1.75	2.72	1.84
0.95	0.78	0.66	0.61	0.89	1.03	1.73	2.70	1.82

Correlations of the estimated discrimination parameter (b^*) with the true discrimination parameters (b_1, b_2, b_3) and their mean (M_b) for dimension 1, 2, and 3. The average absolute differences are AAD_{b1}, AAD_{b2}, AAD_{b3}, and AAD_{Mb}

$\rho(\theta_1, \theta_2, \theta_3)$ are presented in Table 14. With respect to the correlations it can be clearly seen that as $\rho(\theta_1, \theta_2, \theta_3)$ increased there were no appreciable changes in the relationship of the estimated difficulty parameter (b^*) and the true parameters (b_1, b_2, b_3). However, it is clear that the strongest relationship was between b^* and M_b for all levels of $\rho(\theta_1, \theta_2, \theta_3)$

Table 14 also indicates that all AAD's decreased as $\rho(\theta_1, \theta_2, \theta_3)$ increased. Further, the smallest AAD was between b^* and b_1 , indicating that b^* is a better estimate of b_1 . An interesting observation was that b^* consistently overestimated b_1 , but b^* approached b_1 as $\rho(\theta_1, \theta_2, \theta_3)$ increased.

CHAPTER 5

Summary, Conclusions, and Recommendations

Within this study the robustness of UIRT models to the violation of the assumption of unidimensionality was tested. It was clear that when a pseudo two parameter logistic model was tested against a two and three dimensional data set, errors of estimation increased as a function of increased disparity between the dimensions. In other words as the correlation between the dimensions decreased the precision of the estimation procedure was reduced. The results of the current study provide support for outcomes presented by Ansley and Forsyth (1985).

In the two dimensional case, ϕ^* values can best be considered as an average of the true ability values ϕ_1 and ϕ_2 and the strength of this relationship increased as $\rho(\phi_1, \phi_2)$ increased. These results were also true for the three dimensional data structure. The same relationship held for the discrimination parameter when the underlying structure was two dimensional. However, in the three dimensional case the relationship between the estimated discrimination parameters and the true discrimination parameters was found to be far weaker. The estimated discrimination parameters were more highly correlated with the true parameter from the third dimension while the smallest AAD was associated with the average true discrimination parameters. With two dimensional data the difficulty parameter b^* was found to be best represented as an overestimation of the true b_1 ; however, it did approach b_1 as $\rho(\phi_1, \phi_2)$ increased. The same results were found to hold for estimates derived from data generated to have a three dimensional structure. In general it was observed that interpretation of the relationship between estimated parameter and true parameter was far less conclusive for three dimensional data than it was for two dimensional data. The correlations between estimated and true parameters in the three dimensional case were found to be smaller than the correlations between the estimated and true parameters in the two dimensional case. Conversely, the AAD's, on average, were

larger in the data generated to be three dimensional than the AAD's in the data generated to be two dimensional.

UIRT estimation procedures appeared to be robust to violation of the unidimensional assumption when the dimensions were highly related and the dimensionality of the underlying structure was small; however, interpretation of the estimated item and trait parameters is still tenuous. Conversely, as the relationship between the dimensions became more divergent, the robustness of the UIRT model decreased significantly. Further, as the number of dimensions increased interpretability of the estimated unidimensional parameters became untenable.

As in any study, there are a number of limitations, some evident at onset and others are discovered during the process of conducting the research. One such limitation is related to the model employed to generate the multidimensional data sets. The utilization of a noncompensatory multidimensional extension of a unidimensional pseudo two parameter model, to represent response data, has little theoretical support to justify its use. However, given the infancy of such simulations, it must serve as one of the few adequate representation which are currently available. A second limitation, this one discovered in the process, is related to the distribution of the item parameters. The distribution of the item and trait parameters were adjusted to give item statistics similar to that of actual data but perhaps different parameters may have led to other conclusions.

It would be interesting to see if these results would hold for higher dimensions and if they could be replicated if a compensatory model were used for data generation. What is needed are indices which allow for a clearer interpretation of the dimensionality of the data set. Recommendations have been cited which suggest that a form of nonlinear factor analysis could be used, in which one utilizes the absolute sum of squares of residuals and the number of residuals greater than some criterion value set by the researcher. It was hoped in the current study that multidimensional scaling techniques would have been able to recover the underlying structure of the data. However, the results show, at least for the

data generation employed in this study, that for data structures more complex than two dimensions, multidimensional scaling techniques may have little real utility.

To the observer, such an endeavor as the current simulation has an appearance of being a smooth process. One simply looks up in a table of standards the assumptions one is to make regarding the type of model that is to be employed, the form of the distribution for the parameters, and the type of random number generators that will be run. After these are in place one then selects the computer program which will best suit the needs of the project and then generates the response matrix which will allow the researcher to conduct the study. However, no such standards exists, and moreover, most computer programs to generate data are written by the researcher or are at least commissioned by the researcher to suit the purposes of the study. The difficulty of such a situation is that the added variability not only makes conducting such research complex but also makes comparisons of the results practically impossible. In order to reduce the amount of variability due to factors outside the variables of interest, some form of standards for the generation of simulation data should be considered.

REFERENCES

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.
- Andrich, D. (1978). A rating formula for ordered response categories. *Psychometrika*, 43, 123-140.
- Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.
- Ansley, T. N. & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
- Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research*, 47(2), 151-178.
- Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.
- Bartholomew, D. J. (1980). Factor analysis of categorical data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42, 293-321.
- Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22, 231-242.
- Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 17(4), 283-296.
- Birenbaum, M. & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. *Journal of Educational Measurement*, 19(4), 259-266.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Don Mills, ON: Addison Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-52.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of EM algorithm. *Psychometrika*, 46, 443-459.

- Bock, R. D., Mislevy, R. J., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher*, 11, 4-11, 16.
- Coombs, C. H. (1964). *A Theory of Data*. New York: Wiley.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto, ON: Holt, Rinehart and Winston. [chap. 15, pp. 339-375]
- Doody-Bogan, E., & Yen, W. M. (1983). Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic model. Paper presented at the meeting of the American Educational Research Association, Montreal.
- Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
- Goldstein, H. (1980). Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 33, 234-246.
- Hambleton, R. K. (Ed.). (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.
- Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, 10, 287-302.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 48(4), 467-510.
- Hambleton, R. K., & van der Linden, w. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378.
- Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 11, 91-115.
- Hattie, J. (1981). Decision criterion for determining unidimensionality. Doctoral thesis, University of Toronto.

- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 20(1), 49-79.
- Hattie, J. (1985). Methodology review assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2) 139-164.
- Herrnstein, R. J., (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of Experimental Analysis of Behavior*, 4, 267-272.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1983). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6(3), 249-260.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1982). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- International Mathematical and Statistical Libraries, (1987). *IMSL Reference Manual*. Houston TX.
- Iwasa, Y., Higashia, M., & Yamamura, N (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. *The American Naturalist*, 117, 710-723.
- Kruskal, J. B. & Wish, M., (1978). *Multidimensional scaling*. Beverly Hills: Sage Publications, Inc.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517-548.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter Logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Don Mills, ON: Addison Wesley.
- Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin*, 13(2), 122-131.

- Lumsden, J. (1976). Test theory. In M. R. Rosenzweig, & L. W. Porter (Eds.), *Annual Review of Psychology* Palo Alto, CA: Annual Reviews.
- McDonald, R. P. (1981). The dimensionality of a test and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.
- McKinley, R. L., & Reckase, M. D. (1982). *The use of the Rasch model with multidimensional item response data* (Research Report ONR 82-1). Iowa City IA: The American College Testing Program.
- Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, 50(1), 69-82.
- Mislevy, R. J. (1987). *Recent developments in item response theory with implications for teacher certification*. In E. Z. Rothkopf (Ed.), *Review of research in education* (vol. 14) (pp. 239-275). Washington, D.C.: American Educational Research Association.
- Mulaik, S. A. (1972). A mathematical investigation of some multidimensional Rasch models for psychological tests. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.
- Pandey, T. N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 212-229). Vancouver, BC: Educational Research Institute of British Columbia.
- Pearce, J. M. & Hall, G. (1980). Overshadowing the instrumental conditioning of a lever press response by a more valid predictor of the reinforcer. *Journal of Experimental Psychology: Animal Behavioral Processes*, 4, 356-357.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 4, 321-334.
- Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.

- Ree, M. (1979). Estimating item characteristic curves. *Applied Psychological Measurement*, 3, 371-385.
- Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (Eds.), *Classical conditioning, Vol. II: Current theory and research*. New York: Appleton-Century-Crofts.
- Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika*, 52(4), 217-273.
- Ross, J. (1966). An empirical study of a logistic mental test model. *Psychometrika*, 31, 325-340.
- Samejima, F. (1972). A general model for free response data. *Psychometric Monograph* 18.
- Samejima, F. (1974). Normal ogive models on the continuous response level in the multidimensional latent space. *Psychometrika*, 39(1), 111-121.
- Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53-61.
- Statistical Package for the Social Sciences, (1988). *SPSS-X Users Guide* (3rd ed.). Chicago: SPSS Inc..
- Stegelmann, W. (1983). Expanding the Rasch Model to a general model having more than one dimension. *Psychometrika*, 48(2), 259-268.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589-617.
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.
- Takane, Y., Young, F. W., & Leeuw, J., (1977). Nonmetric individual multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1), 7-67.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49(1), 95-110.

- Tatsuoka, M. M. (1968). Mathematical models in the behavioral and social sciences. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences*, 3-59. Addison-Wesley, Don Mills, Ont..
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 51(2), 201-214.
- Traub, R. E. (1983). A priori assumptions in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.
- Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of Research in Education* (vol. 9) (pp. 377-435). Washington, D.C.: American Educational Research Association.
- Tuker, L. R., Humphreys, L. G., & Roznowski, M. A. (1986). Comparative accuracy of five indices of dimensionality of binary items. Champaign-Urbana, IL: University of Illinois, Department of Psychology.
- Warm, T. A. (1978). *A primer of item response theory*. Springfield, VA: U.S. Department of Commerce, National Technical Information Service.
- Way, W. D., Ansley, T. N. & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT models. *Applied Psychological Measurement*, 12, 239-252.
- Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45(4), 479-494.
- Whitely, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement*, 18(2), 67-84.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). Logist user's guide. Princeton, NJ: Educational Testing Service.
- Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45-56). Vancouver, BC: Educational Research Institute of British Columbia.

- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97-116.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
- Yen, W. M. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 123-141). Vancouver, BC: Educational Research Institute of British Columbia.